

**Butler, Jennie C**

**From:** Wayne R. Kubick [wayne-kubick@prosys-llc.com]  
**Sent:** Tuesday, November 02, 1999 4:18 PM  
**To:** FDADOCKETS@OC.FDA.GOV  
**Subject:** Comments on Dockets 97D-0381 and 98D-0317 Regarding Electronic Regulatory Submissions

On behalf of the Clinical Data Interchange Standards Committee (CDISC), we wish to submit the attached document as a comment to FDA Docket numbers 97D-0381 and 98D-0317 related to the CDER/CBER Guidance for Industry: Providing Regulatory Submissions in Electronic Format - General Considerations (January 1999 IT 2).

This document describes a proposed metadata model that can be applied to electronic datasets that are supplied to CDER and CBER as SAS V5 Transport files. By applying the CDISC Metadata model to these submissions, industry sponsors will:

- Provide FDA reviewers with clear descriptions of the usage, structure, contents and attributes of all datasets and variables
- Allow FDA reviewers to replicate most analyses, tables, graphs and listings with minimal or no transformations
- Enable FDA reviewers to easily view and subset the data used to generate any analysis, table, graph or listing without complex programming.

We believe the FDA should consider this model as they prepare future versions of the Guidance and look forward to further enhancing and developing the model as a way to facilitate the process of data standardization within the pharmaceutical and biotechnology industry.

Thank you for considering this comment.

David Christiansen	Wayne R. Kubick
Genentech	PROsys LLC
davec@gene.com	wayne-kubick@prosys-llc.com
650-225 1738	847 842-1846



Submission Metadata  
Model1.pdf...

## CDISC Submission Metadata Model

Version 0.21 – 2 November 1999

### The CDISC Submission Data Standards Group Metadata Approach

The CDISC Submission Data Standards Group has chosen to focus on metadata as a more attainable way of establishing meaningful standards applicable to data submitted for FDA review. Metadata is defined as “data about the data”; in other words, metadata includes description of the content, context, structure and/or purpose of a database. It is important to recognize that the metadata provided by the model is intended to be the minimum required to meet the need of FDA users and is not intended to fully meet all of the needs of the sponsor’s data management, statistics or other internal groups. Additional internal metadata standards will be desirable within most organizations to govern the ways that data is captured, cleaned and analyzed statistically.

Submission datasets are described in the FDA “Guidance for Industry, **Providing Regulatory Submissions in Electronic Format — NDAs** (IT 3 January 1999) under Item 11: Case Report Tabulations (CRTs). The Guidance states:

“Each dataset is a single file and, in general, includes a combination of raw and derived data. Just as you provide each CRF domain (e.g., demographics, vital signs, adverse events) as a table in a paper submission, in an electronic submission, each CRF domain should be provided as a single dataset. In addition, datasets suitable for reproducing and confirming analyses may also be needed. Prior to the submission, you should discuss with the review division the datasets to be provided and the data elements that should be included in each dataset.”

The CDISC Submission Metadata Model was created to help ensure that the supporting metadata for these submission datasets should meet the following objectives:

- Provide FDA reviewers with clear descriptions of the usage, structure, contents and attributes of all datasets and variables
- Allow reviewers to replicate most analyses, tables, graphs and listings with minimal or no transformations
- Enable reviewers to easily view and subset the data used to generate any analysis, table, graph or listing without complex programming.

The model does not address specific content issues such as how to populate individual tables for a particular study. The data collected and reported for a study should be based on scientific and medical considerations. However, the model does guide sponsors toward certain common conventions that should provide greater consistency and uniformity among all future submissions and begin to reduce the range of possibility for diversity in data that is provided for regulatory submissions. The model helps ensure that those data domains, elements and attributes that are common to all submissions will be represented in the same manner in every case. By concentrating on metadata rather than actual data structures, the Submissions Group hopes to achieve many of the benefits of dictionary standardization while ensuring that scientific objectives are not compromised.

In order to achieve their goals, the Submissions Group has been working from information supplied by the FDA. The role of the Submissions Group is to suggest an approach to FDA; but this approach is not to be taken as a requirement for implementation until it is incorporated into FDA guidance documents or regulations. The CDISC Submission Data Standards are based on the FDA and the Example of an Electronic New Drug Application

Submission dated 2/17/99. See [Summary of Metadata Model Changes](#) for a history of proposed changes for these documents.

## Submitting Comments

All comments should be submitted by using the message board facility provided on the CDISC Data Submissions Group web page at <http://www.diahome.org/cdisc>.

## Metadata Definitions for Domain Datasets

In the guidance, the core safety data is divided into 12 domain datasets and the data elements are described in the data definition file (define.pdf). Using the FDA Sample NDA as a starting point, the CDISC Submission Data Standards Group has augmented the FDA sample and proposes the following dataset definitions.

The first table in the data definition file provides basic information about each of the datasets in four columns:

- **DATASET NAME** - The 8 character name of the dataset is provided by the sponsor in this column. Note: the FDA may be defining preferred names to be used for some of the more common datasets at a later date. For this example, the names used in the Sample NDA will be used whenever possible.
- **DESCRIPTION** - A more detailed description of the information contained in the dataset is included in this column. This column should also be linked to the domain variable descriptions table.
- **LOCATION** - The file location including the folder and file name is included in this column. This column may be linked to the dataset.
- **STRUCTURE** - The level of detail represented by each record in a dataset. In other words what is the “shape” of the dataset? Examples of dataset structure include one record per patient, one record per patient per visit, one record per patient per event, one record per patient per visit per event, etc. A single data domain may need more than one structure to facilitate understanding of the data (for example, labs may be reported as one record per lab measurement, or as one record per patient visit). The structure(s) for a dataset may depend on the type of data, the indication, and/or reviewer preferences. This information was not described in the guidance.

## Dataset Definition

Description of datasets			
Dataset	Description	Location	Structure
<a href="#">DEMO</a>	Demographics and Subject Characteristics	Demo.xpt	1 Rec/patient
<a href="#">CONMEDS</a>	Concomitant medication	Conmed.xpt	1 Rec/patient/Incident
<a href="#">EXPOSE</a>	Drug exposure	Expose.xpt	
<a href="#">DISPOSIT</a>	Disposition	Disposit.xpt	
<a href="#">AE</a>	Adverse Events	Adv.xpt	1 Rec/patient/adverse event
<a href="#">CHEM</a>	Labs – chemistry detail	Chem.xpt	1 Rec/patient/visit/measurement
<a href="#">CHEMSUM</a>	Labs – chemistry summary	Chemsum.xpt	1 Rec/patient/visit
<a href="#">HEMAT</a>	Labs – hematology	Hemat.xpt	
<a href="#">URINE</a>	Labs – urinalysis	Urine.xpt	

<a href="#">ECG</a>	ECG	Ecg.xpt	
<a href="#">VITAL</a>	Vital signs	Vital.xpt	
<a href="#">PE</a>	Physical examination	Pe.xpt	
<a href="#">MEDHIST</a>	Past medical history	Medhist.xpt	

## Metadata Definitions for Domain Variables

These tables describe the specific variables included in each dataset. The CDISC metadata model intends to create a superset of possible variables that might be included in a submission over time – not a standard list of required elements. Only the FDA can specify required variables -- which would be defined in the form of future guidance documents. While there are some data elements that would normally be expected to exist in the datasets for most submissions, there are many others that are indication-specific or optional and should only be included when pertinent to the submission at hand. As a general principal, a data submission should only include those variables that are relevant to the analysis of safety and efficacy performed for that study.

The metadata definitions for each domain include the following metadata columns:

**SUGGESTED VARIABLE NAME** - This column should include the 8-character field name the sponsor used for its analyses. The 8-character limitation is currently required due to a limitation of the SAS V5 transport format currently required by the FDA; a more flexible format is expected to replace this in the near future. (Note: FDA will be defining standard field names that should be used by sponsors where possible once they are made available.) Early versions of the model list standard variable names for certain key and selection variables only. It is assumed that the sponsor will provide names for other variables using their internal variable names. Once a common standard for naming other variables included in a submission is eventually be defined by the FDA in a future guidance document, the sponsor's internal names can still be included as aliases.

**VARIABLE LABEL** - This is a 40-character description of the variable (the maximum length allowed by SAS V5 transport datasets). The label should adequately explain the content of the variable.

**TYPE** - This describes if the variable is a character string or numeric value to conform with existing SAS conventions. Values that can be either character or numeric are listed as “Char or Num”, but should be consistent throughout a submission. Since date, time and date-time values are stored by SAS as numeric values, you should specify SAS formats in the DECODE/FORMAT column where appropriate.

**DECODES/FORMATS** - This column describes the character or number codes used in the dataset for each variable. Wherever possible, text should be used instead of codes (e.g., use “Yes” instead of “1”). In addition, this column can list all allowable values for the field. (This column could contain standard SAS formats)

**ORIGIN** - This column shows the point of origin for each variable included within the current domain. .

- **Current Domain** - For variables that originate within the current domain, the column should indicate whether it was collected intact from a CRF or electronic device (a source variable) or whether it was computed or derived.
  - **Source Variables** - For variables captured from CRFs, the location of the information in the case report form is provided as a link to the annotated CRF pdf file.

- **Derived Variables** - For derived variables, a description of how the variable was derived including any algorithms used is provided (this may be done by providing a hyperlink to another document describing the derivation algorithm or process especially when the algorithm is complex).
- **External Domain** - For variables that originate in other, external domains but are merged into the current domain (such as to help the reviewer more easily subset or sort the data), this column should list the dataset where it originated and provide a link to the appropriate domain variable definition. For example, the variable “Age” would be described in the AE variable definition table as Demog.Age; and a hyperlink to the DEMOG definition table would then show how age was computed in the Demog Domain.

**ROLE** - This information provides information on how a variable is used in the analysis of the associated dataset. In some cases, more than one role may be listed for a specific variable. There are four types of Roles for variables:

- **Key Variables** - used to uniquely identify and index each record in a dataset: Study, Center, Patient ID, Visit, Event Nr. Most datasets will have between 3 and 5 key variables. Key variables are often used to link tables when merging in additional fields. Key Variables may also be used for grouping, sorting or selecting subsets of data. Key variables should **always** be clearly identified in the metadata – all other roles are optional but should be defined where possible to help the reviewer understand why the data was included.
- **Selection Variables** – fields that are not Keys but which are frequently used to subset, sort or group data for reporting purposes. These fields are frequently included into other datasets to simplify queries and facilitate simple analyses. For example: Sex, Age and Race are often merged into other datasets from Demographics, and the FDA requests that Treatment Group be merged from the Drug Exposure table into every submitted domain.
- **Review Variables** – fields that are not Key or Selection variables, but are source or derived variables related to the objectives of the study. These variables are usually tabulated and or summarized for analysis purposes. They may be continuous, categorical or both (e.g., mean age and frequency by age group).
- **Support Variables**– these variables are not Key, Selection or Review variables, but provide other useful background or reference information, or provide input for deriving Review variables.
- **COMMENTS** – This field is used to present other useful information that may assist the reviewer in understanding the data. For example, it might list which coding dictionary (COSTART, MEDDRA, etc.) was used to populate a coded value.

## Data Management Conventions for Populating Datasets

The CDISC Metadata Model describes the structure and form of data, not the content. However, the varying nature of clinical data in general will require the sponsor to make some decisions about how to represent certain real world conditions in the dataset. Therefore, it is useful for a metadata document to give the reviewer an indication of how the datasets handle certain special cases. Some of these special cases to be considered and suggestions for handling them are described below:

- Case sensitivity of Text values -- The metadata document should indicate if uppercase is used consistently for text data.
- Missing values – Since data will be submitted as SAS transport datasets, the convention used for missing values should be described. The conventions should be used consistently in all datasets and explained in the metadata.

- Non-numeric data in numeric fields – In general an alphanumeric field type must be used for a numeric field that contains non-numeric data, but it is useful if the metadata describes when this occurs.
- Partial dates - the metadata should indicate whether dummy dates or incomplete values are used for specified dates, and the same convention should be used consistently.
- Partial date/times - these should be handled in the same manner as partial dates.

Additional conventions are expected to be included in future versions of the submission metadata model.

## Metadata for a Generic Domain

The following generic domain applies the metadata definitions for domain variables and can be used as a guide in preparing a metadata table for any submission dataset.

### Domain Name

Suggested Variable Name (8 Char)	Variable label (40 Character)	Type	Decodes/ Formats	Origin	Role	Comments
STUDYID	Study identification	Char		<a href="#">Domain CRF Page</a>	Key	
SITEID	Center or site identification	Char		<a href="#">Domain CRF Page</a>	Key	
PID	Patient identification	Char or Num		<a href="#">Domain CRF Page</a>	Key	
	Other Key Value (Event, Item, etc.)	Char or Num		<a href="#">Domain CRF Page</a>	Key	
	Patient's initials	Char		<a href="#">Domain CRF Page</a>	Support	
INVNAME	Investigator name	Char		Site.Investigator	Support	
COUNTRY	Country	Char		Demog.	Support	
BIRTHDT	Birth date	Num	SAS Date	Demog.	Support	
VISITNUM	Visit	Num		<a href="#">Domain CRF Page</a>	Sel	
VISITDT	Visit Date	Num	SAS Date	<a href="#">Domain CRF Page</a>		
	Study Day	Num		Domain.Visit-Date – Demog.Baseline-Date	Sel	
AGE	Age in years at baseline	Num		Demog.	Sel	Depends on Protocol definition
SEX	Gender/Sex	Char	female, male	Demog.	Sel	
RACE	Race	Char	Caucasian, Black, Asian, Hispanic, Other	Demog.	Sel	
TRTNAME	Treatment name	Char		Exposure.Treat Name	Sel	
TRTCODE	Treatment code	Num		Exposure.Treat Code	Sel	

## Metadata Examples for Clinical Safety Domains

The metadata model provides metadata examples for the twelve important clinical safety domains specified as high priorities in the FDA guidance document. These tables are based on the information provided in the guidance document but do not provide an official FDA opinion of the data elements. CDISC has prepared example metadata

models for each of these domains consistent with the guidelines presented in this document. For these examples, the variable names are specified only for key and common selection variables that have been suggested by the FDA.

[Clinical Domain Data Models](#)

[CDISC Submissions Data Standards Group Web Page](#)